AD_____

Award Number: DAMD17-98-2-8005

TITLE: Malaria Genome Sequencing Project

PRINCIPAL INVESTIGATOR: Malcolm J. Gardner, Ph.D.

CONTRACTING ORGANIZATION: The Institute for Genomic Research
Rockville, Maryland 20850

REPORT DATE: January 2002

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	January 2002	Annual (17 Dec 00 - 16 Dec 01)

4. TITLE AND SUBTITLE
Malaria Genome Sequencing Project

5. FUNDING NUMBERS
DAMD17-98-2-8005

6. AUTHOR(S)
Malcolm J. Gardner, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Institute for Genomic Research
Rockville, Maryland 20850

E-Mail: gardner@tigr.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The objectives of this 5-year Cooperative Agreement between TIGR and the Malaria Program, NMRC, were to: **Specific Aim 1**, sequence 3.5 Mb of *P. falciparum* genomic DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. Two additional Specific Aims were added to the Cooperative Agreement: **Specific Aim 4**, sequencing of *P. yoelii* to 3X coverage; **Specific Aim 5**, sequencing of *P. vivax* to 3X coverage. To date, we have published the first complete sequence of a malarial chromosome (*P. falciparum* chromosome 2), have sequenced and have nearly completed the sequences of *P. falciparum* chromosomes 10, 11, and 14. A plan for joint annotation and publication of the *P. falciparum* genome with the Sanger Centre and Stanford University was prepared and implemented. Publication of the *P. falciparum* genome sequence is anticipated in late 2002. In addition, we completed sequencing of the rodent malaria *P. yoelii* to 5X coverage and have begun the sequencing of *P. vivax* to 5X coverage. Discussions with the Malaria Program, NMRC aimed at development of a program to use genomics and functional genomics to accelerate vaccine research are in progress.

14. SUBJECT TERMS
p. falciparum, P. vivax, P. yoelii, malaria, genome, chromosome

15. NUMBER OF PAGES
25

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Introduction

Malaria is caused by apicomplexan parasites of the genus *Plasmodium*. It is a major public health problem in many tropical areas of the world, and also affects many individuals and military forces that visit these areas. In 1994 the World Health Organization estimated that there were 300-500 million cases and up to 2.7 million deaths caused by malaria each year, and because of increased parasite resistance to chloroquine and other antimalarials the situation is expected to worsen considerably. These dire facts have stimulated efforts to develop an international, coordinated strategy for malaria research and control (3). Development of new drugs and vaccines against malaria will undoubtedly be an important factor in control of the disease. However, despite recent progress, drug and vaccine development has been a slow and difficult process, hampered by the complex life cycle of the parasite, a limited number of drug and vaccine targets, and our incomplete understanding of parasite biology and host-parasite interactions.

The advent of microbial genomics, i.e. the ability to sequence and study the entire genomes of microbes, should accelerate the process of drug and vaccine development for microbial pathogens. As pointed out by Bloom, the complete genome sequence provides the "sequence of every virulence determinant, every protein antigen, and every drug target" in an organism (2), and establishes an excellent starting point for this process. In 1995, an international consortium including the National Institutes of Health, the Wellcome Trust, the Burroughs Wellcome Fund, and the US Department of Defense was formed (Malaria Genome Sequencing Project) to finance and coordinate genome sequencing of the human malaria parasite *Plasmodium falciparum*, and later, a second, yet to be determined, species of *Plasmodium*. Another major goal of the consortium was to foster close collaboration between members of the consortium and other agencies such as the World Health Organization, so that the knowledge generated by the Project could be rapidly applied to basic research and antimalarial drug and vaccine development programs worldwide.

Body

This report describes progress in the Malaria Genome Sequencing Project achieved by The Institute for Genomic Research and the Malaria Program, Naval Medical Research Center, under Cooperative Research Agreement DAMD17-98-2-8005, over the 12 month period from Dec. '00 to Dec '01. The Specific Aims of the work supported by this agreement are listed below. Specific Aims 1-3 were contained in the original Cooperative Agreement. Specific Aims 4-5 were added to the Cooperative Agreement through modifications.

1. Determine the sequence of 3.5 megabases of the *P. falciparum* genome (clone 3D7):

a) Construct small-insert shotgun libraries (1-2 kb inserts) of chromosomal DNA isolated from preparative pulsed-field gels.

b) Sequence a sufficiently large number of randomly selected clones from a shotgun library to provide 10-fold coverage of the selected chromosome.

c) Construct P1 artificial chromosome (PAC) libraries (inserts up to 20 kb) of chromosomal DNA isolated from preparative pulsed-field gels.

d) If necessary, generate additional STS markers for the chromosome by i) mapping unique-sequence contigs derived from assembly of the random sequences to chromosome, ii) mapping end-sequences from chromosome-specific PAC clones to YACs.

e) Use TIGR Assembler to assemble random sequence fragments, and order contigs by comparison to the STS markers on each chromosome.

f) Close any remaining gaps in the chromosome sequence by PCR and primer-walking using *P. falciparum* genomic DNA or the YAC, BAC, or PAC clones from each chromosome as templates.

2. Analyze and annotate the genome sequence:

a) employ a variety of computer techniques to predict gene structures and relate them to known proteins by similarity searches against databases; identify untranslated features such as tRNA genes, rRNA genes, insertion sequences and repetitive elements; determine potential regulatory sequences and ribosome binding sites; use these data to identify metabolic pathways in *P. falciparum.*

3. Establish a publicly-accessible *P. falciparum* genome database and submit sequences to GenBank.

4. Perform whole genome shotgun sequencing of the rodent malaria parasite *Plasmodium yoelii* to 3X coverage, assemble into contigs, annotate the contigs, make the data available on the TIGR web site, and submit the data to GenBank.

5. Perform whole genome shotgun sequencing of the human malaria parasite *Plasmodium vivax* to 3X coverage, assemble the contigs, annotate the contigs, make the data available on the TIGR web site, and submit the data to GenBank.

We are pleased to report that excellent progress has been made towards achievement of these goals. In previous annual reports we announced the publication in *Science* of the first complete sequence of a malarial chromosome (chromosome 2) (5); development of a *Plasmodium* gene finding program, GlimmerM (9); introduction of optical restriction mapping technology for rapid mapping of whole *Plasmodium* chromosomes (6, 8); completion of the random phase of sequencing of 3 additional *P. falciparum* chromosomes and major progress in gap closure; use of microarray technology to examine the expression of all genes from chromosomes 2 and 3 of *Plasmodium* (by our collaborators at the Naval Medical Research Center); construction of a *P. falciparum* genome web site at TIGR which contains all of the

chromosome 2 sequence data and annotation, as well as preliminary sequences for the 3 other chromosomes currently being sequenced (http://www.tigr.org/tdb/mdb/pfdb/pfdb.html); sequencing of the rodent malaria parasite *Plasmodium yoelii* to 3X coverage and release of preliminary annotation of this genome on the TIGR web site (http://www.tigr.org/tdb/edb2/pya1/htmls/). Through a subcontract to Dr. John Yates at the Scripps Institute, we also assisted NMRC in their pilot project to apply the techniques of proteomics towards the identification of novel antigens in parasite (sporozoite) extracts. Finally, we have continually reviewed with NMRC further steps that can be taken to more rapidly apply *Plasmodium* genomics, functional genomics, and proteomics to problems of vaccine development for malaria.

Over the past year, efforts have focused on gap closure of *P. falciparum* chromosomes 10, 11, and 14. In addition, after extensive discussions with the other members of the malaria genome consortium, we have devised and implemented a plan for the joint annotation and publication of the *P. falciparum* genome with the Sanger Centre and Stanford University. Finally, we completed the sequencing of *P. yoelii* to 5X coverage, and initiated work on the sequencing of the *P. vivax* genome.

Sequencing of *P. falciparum* chromosomes 10, 11, and 14 (Specific Aims 1, 2, 3)

Sequencing of chromosome 10, 11, and 14 is being funded primarily by grants from the NIAID (chromosomes 10 and 11) and the Burroughs Wellcome Fund (chromosome 14). Funds from this collaborative agreement are being used to accelerate the sequencing, assist in closure and annotation, develop microarrays for chromosome 14, and facilitate rapid utilization of the sequence data by the DoD vaccine and drug development groups. In previous years we described the isolation of chromosome 14 DNA, preparation of shotgun libraries , random sequencing, assembly, progress in gap closure, production and public release of preliminary annotation. This past year focused primarily on gap closure and preparations for the final annotation of the whole *P. falciparum* genome.

The current status of closure for chromosomes 10, 11, and 14 is indicated in Table 1. All of the gaps that remain are in regions that are difficult to clone and or sequence, but none of the gaps are expected to be longer than 1 kb. We expect that most of these gaps will have been closed by the time that the genome has been published and this cooperative agreement expires in Dec. 2002.

Table 1. Status of closure for chromosomes 10, 11, and 14

Chromosome	Length (Mb)	Sequence gaps	Physical gaps
10	1.69	0	4
11	2.04	0	3
14	3.29	1	0

Annotation and publication of the *P. falciparum* genome sequence (Specific Aims 2 and 3)

In last year's report we described the release of preliminary sequence data and annotation for chromosomes 10, 11, and 14 on the TIGR web site, and outlined the basic procedures that would be used to annotate these chromosomes. Our original plan was to annotate chromosomes 10, 11, and 14 and publish them together in a single article. A major drawback to this plan was that publication of this small fraction of the *P. falciparum* genome would not allow the analysis of the genome in its entirety, and hence fundamental questions of organism biology, biochemistry, and pathogenesis could not be addressed.

Discussions with our counterparts at the Sanger Centre and Stanford University over the past year have led to an agreement for the 3 centers to collaborate on a joint publication of an analysis of the entire *P. falciparum* genome sequence. This whole genome overview will be accompanied by a series of papers by each sequencing center on the chromosomes sequenced by each group. The whole genome overview and chromosome papers will be published in a single issue of a journal. In addition, a comparative analysis of the *P. falciparum* and *P. yoelii* genomes based upon the 5X coverage *P. yoelii* sequence will be published along with the P. *falciparum* papers.

The basic elements of this plan (4; Appendix 1) include beginning the annotation on a set of contig sequences representing the best available data for each chromosome. These contigs would be "frozen" so to permit annotation to proceed on a stable data set, and where possible the contigs will be joined end-to-end in the correct order and orientation to form draft chromosome sequences. As the annotation of these draft chromosomes is underway, closure efforts on the remaining gaps will continue, and the new sequence data generated during the closure process will be merged into the annotated contigs near the end of the process. Each sequencing center will be responsible for annotation of the chromosomes they sequenced, using the software and methods in use at each center. In an attempt to ensure that the annotation done by the participating centers is of equal quality, the same 100 kb sequence will be annotated by the three groups early in the annotation process and the results will be compared to identify any problems. Furthermore, it was agreed that TIGR will maintain a central relational database containing a representation of the sequence data and annotation produced at all three centers, and that the centers will develop procedures for the frequent semi-automated exchanges of data. This will allow all of the annotators to view the same picture of the complete genome and facilitate whole genome analyses. Importantly, this arrangement will also simplify the process of submitting the annotated genome sequence to the PlasmoDB database (1). This plan has now been put in motion. Many chromosome sequences have been frozen, annotation has begun, and the system for data exchange between centers is being tested.

Sequencing of *P. yoelii* and *P. vivax* to 3X coverage (Specific Aims 4, 5)

A secondary goal of the malaria genome project was to sequence the genome of a another species of *Plasmodium*, and discussions as to which parasite should be chosen had generated lively discussions amongst the malaria community, with some groups favoring sequencing of the

human malaria *P. vivax*, and others advocating sequencing one of the rodent malaria parasites that are used as model systems. The sequence of one or more species would be very useful for comparison to *P. falciparum*, perhaps enabling the identification of genes that may be involved in differences in life cycles and pathogenicity, for example. Genome sequence information from other *Plasmodium* species would also be helpful in annotation of *P. falciparum*, by assisting in identification of genes conserved across different species. Recent discussions at the semi-annual meetings of the malaria genome consortium may lead to efforts funded by the NIAID, the Burroughs Wellcome Fund, or the Wellcome Trust, to do partial sequencing of several rodent malaria genomes, which will provide useful sequence data to groups working on these different parasites at a reasonable cost.

In light of these events, and the reductions in sequencing costs achieved by the TIGR SeqCore through improvements in instrumentation and sequencing protocols, the TIGR/NMRC team discussed the expansion of our sequencing efforts to include *P. vivax* and *P. yoelii*. *P. vivax* is a major human malaria parasite, and *P. yoelii* is a rodent malaria parasite used as a model system for vaccine development by NMRC. By using a whole genome shotgun strategy and sequencing to 3X coverage , it is possible to assemble contigs covering about 90% of the *Plasmodium* genome. With the high gene density of *Plasmodium*, this is a relatively rapid and low-cost method to acquire partial or complete sequences of almost all parasite genes.

After discussions with NMRC we elected to proceed with sequencing of *P. yoelii* and then to sequence *P. vivax* These projects were initially under the supervision of Dr. Leda Cummings, but Dr. Cummings left TIGR in 2001 and the work has been taken over by Dr. Jane Carlton. *P. yoelii* was given priority over *P. vivax* for sequencing despite the greater importance of *P. vivax* as a human pathogen because we did not have *P. vivax* genomic DNA suitable for sequencing (but see below).

Whole genome shotgun (WGS) sequencing of *Plasmodium yoelii yoelii* strain 17XL was completed this year to a level of 5X coverage of the genome. 223,907 sequences of average length 661 bp were assembled into contigs using the TIGR Assembler algorithm. An apparent change in the GC content between contigs greater than 2kb in size and those less than 2kb identified mouse host contamination as a significant problem. Contaminating host mouse sequences were identified through BLASTN comparison of the contigs and singletons with a custom-made mouse/rat database, and excluded from further analysis. Approx. 7,000 contigs with a total cumulative length of 25Mb and average length of 3.5kb are now frozen. Statistics regarding the contigs and singletons are shown in Table 2.

Table 2. Shotgun sequencing of *P. yoelii* to 5X coverage.

	Number contigs	Cumulative length	Average length	Average redundancy	% G+C with mouse	% G+C no mouse
All contigs	**7,302**	**25 Mb**	**3.5 kb**	**3 X**		
All singletons	**24,942**	**15 Mb**	**615 bp**	-	**39.3**	**32.8**
Contigs > 1 kb	5689	24 Mb	4.2 kb	3.4 X	32.7	31.8
Contigs > 2 kb	2979	20 Mb	6.8 kb	5.2 X	23.2	23.1
Contigs > 5 kb	1440	15 Mb	10.6 kb	6.0 X	22.3	22.3
Contigs > 10 kb	571	9 Mb	16.0 kb	6.5 X	22.5	22.5
Contigs > 20 kb	111	2 Mb	26.5 kb	6.4 X	22.8	22.7
Contigs > 30 kb	21	812,178 kb	38.7 kb	6.8 X	23.2	23.2
Contigs > 40 kb	8	376,885 kb	47 kb	7.3 X	23.4	23.4
Contigs > 50 kb	2	102,781 kb	51.4 kb	7.7 X	23.7	23.7
Contigs < 1kb + singletons	**26552**	**16.6 Mb**	**624 bp**	**1.0 X**	**40.7**	**36.4**

All *P. yoelii* genome data has now been released on TIGR's ftp site at http://www.tigr.org/tdb/edb2/pya1/htmls/ for use by the malaria community. A new exon-finder, GlimmerMExon, has been designed and used in conjunction with two other gene finding algorithms to predict genes in all contigs > 2kb. A total of 5,924 complete and 1,825 incomplete gene models have been identified among the contigs. This final set of gene models has been frozen and is currently proceeding through automated gene annotation.

Work is proceeding towards publication of the *P. yoelii* WGS project in conjunction with the *P. falciparum* whole genome analysis for publication in a single journal in the latter part of this year. Although interesting in its own right as a species of malaria, the importance of the *P. yoelii* genome data will be in providing insights into the biology of the human malaria parasite *P. falciparum* through comparative genomics. Comparative analyses of the two genomes, such as identification of orthologous genes, is currently underway. A tiling path of *P. yoelii* contigs along each *P. falciparum* chromosome has been constructed using the PROMer algorithm, highlighting the large degree of synteny between the two species. PCR between the *P. yoelii* contigs using the *P. falciparum* genome as a scaffold has identified regions of the genomes where synteny breaks. These regions are currently being studied for common motifs and other signature sequences.

In the final year of this project we will sequence the SalI strain of *P. vivax* to 5X coverage. Genomic DNA was kindly provided by Dr. John Barnwell of the Centers for Disease Control and two genomic shotgun libraries with inserts sizes ranging from 2-3 kb and 5-6 kb have been prepared. About 100 test sequences have been generated from each library. An analysis of these sequences suggests that the libraries are of sufficient quality for further sequencing. An additional 2000 sequences will be generated from each library and assembled to

determine whether these libraries are random . If the assembly confirms that the shotgun libraries are random, we will proceed with high-throughput sequencing to achieve 5X coverage of *P. vivax* genome.

In addition to the genomic sequencing efforts, we also generated expressed sequence tags (ESTs) from a *P. yoelii* sporozoite cDNA library provided by Dr. Victor Nussenzweig at New York University School of Medicine. The ESTs were submitted to Genbank (accession numbers AF390551-AF390553) and used to construct a gene index that is available on the TIGR web site (http://www.tigr.org/tdb/pygi/). This work was published in the *Proceedings of the National Academy of Sciences (7; Appendix 2).*

To summarize the sequencing efforts in Specific Aims 1-5, by the end of this cooperative agreement, the complete genome sequence of *P. falciparum* will have been published, as will a comparative analysis of the *P. falciparum* and *P. yoelii* genomes. The *P. vivax* genome will have been sequenced to 5X coverage. A 3-way comparative analysis of the *P. falciparum, P. vivax,* and *P. yoelii* genomes will be also be performed, but this is unlikely to be completed until after this cooperative agreement has expired.

Proteomics studies

A major goal of the malaria genome project is to identify antigens for vaccine development. Analysis of the genome sequence data can be used to identify potential antigens but does not by itself provide all of the information required for selection and prioritization of vaccine candidates. For example, the genome sequence itself does not specify at which point in the life cycle a gene is transcribed, or whether the protein product of a gene is actually present in the parasite. To gather information on gene expression patterns we initiated the microarray studies in collaboration with NMRC that are described above. To identify proteins present in various stages of the parasite life cycle, we have begun to use proteomics techniques to directly identify parasite proteins in cell lysates.

Last year we reported on work done by Dr. John Yates at the Scripps Research Institute, partly funded by a subcontract from TIGR under this cooperative agreement. Briefly, proteins in parasite lysates were digested with proteases and the resulting peptides were separated by high resolution liquid chromatography. The peptides were then injected into a tandem mass spectrometer. Spectra of each peptide were matched against predicted spectra of the peptides predicted from the genome sequence. In this way peptides generated from cell lysates were used to identify the proteins present in the cell lysate. Dr. Yates performed a series of such experiments using sporozoites of *P. falciparum* and *P. yoelii* and identified 308 unique proteins from *P. falciparum*, and 37 unique proteins from *P. yoelii*. This represented a massive increase in the number of known sporozoite proteins, and indicates that the same techniques can be used to identify proteins present in other stages of the life cycle. When combined with information gleaned from the sequence data, such as predicted subcellular location, hydophillicity, and predicted T cell epitopes, the protein expression data will help to prioritize potential antigens for vaccine development.

Dr. Yates has continued this work using other funds, in close collaboration with the Malaria Program, NMRC. This year, our role has been to provide his laboratory with preliminary genome sequences, and with predicted protein sequences, from the *P. falciparum* and *P. yoelii* genomes.

Key Research Accomplishments

1) Closure of chromosomes 10, 11, and 14 of *Plasmodium falciparum* was the major focus of work over the past year. All 3 chromosomes are now in the late stages of gap closure. Completion and annotation of these chromosomes is expected by June, 2002.

2) A plan for annotation and publication of the *P. falciparum* genome was developed in consultation with the other members of the Consortium, the Sanger Centre and Stanford University. Publication of the *P. falciparum* genome sequence is expected in late 2002.

3) Shotgun sequencing of the rodent malaria parasite *P. yoelii* to 5X coverage was completed; preliminary contigs and annotation were released on a new TIGR web site http://www.tigr.org/tdb/edb2/pya1/htmls/). Annotation of the preliminary sequence, and a comparative analysis with *P. falciparum* is underway.

4) ESTs were generated from a *P. yoelii* sporozoite cDNA library and used to build a P. yoelii gene index ((7)).

5) Plans to shotgun sequence the *P. vivax* genome to 5X coverage were prepared in consultation with NMRC. Genomic DNA was recently provided to us by Dr. John Barnwell of the CDC and two genomic shotgun libraries prepared. Sequencing to 5X coverage has begun and will be completed by the end of this project.

Reportable Outcomes

1) Web site. Preliminary contigs and annotation for *P. yoelii* genome at 5X coverage. (http://www.tigr.org/tdb/edb2/pya1/htmls/).

2) Web site. *P. yoelii* gene index (http://www.tigr.org/tdb/pygi/).

3) Publication. S. H. Kappe *et al.*, Exploring the transcriptome of the malaria sporozoite stage, *Proc Natl Acad Sci U S A* **98**, 9895-900 (Aug 14, 2001).

4) *P. yoelii* sporozoite ESTs submitted to GenBank (accession numbers AF390551-AF390553.

5) Publication. M. J. Gardner, A status report on the sequencing and annotation of the *P. falciparum* genome, *Mol Biochem Parasitol* **118**, 133-8 (Dec, 2001).

6) M. J. Gardner, Update on sequencing of *P. falciparum* chromosomes 10, 11, and 14. Paper presented at the 10th Meeting of the Malaria Genome Sequencing Consortium, Philadelphia, PA, Feb. 2-4, 2001.

7) M. J. Gardner, Update on plans for *P. falciparum* publication. Paper presented at the 11th Meeting of the Malaria Genome Sequencing Consortium, Hinxton, England, June 5-6, 2001.

8) M. J. Gardner, Sequencing the genome of *P. falciparum*. Paper presented at the 36th Joint Conference on Parasitic Diseases, National Institutes of Health, Bethesda 2001.

9) *Plasmodium falciparum* Annotation Meeting, The Institute for Genomic Research, Rockville, MD, Dec. 6-7, 2001. A meeting of the malaria genome consortium organized by the Principal Investigator to finalize plans for the annotation and publication of the *P. falciparum* genome.

Conclusions

The objectives of this 5-year Cooperative Agreement between TIGR and the Malaria Program, NMRC, were to: **Specific Aim 1**, sequence 3.5 Mb of *P. falciparum* genomic DNA; **Specific Aim 2**, annotate the sequence; **Specific Aim 3**, release the information to the scientific community. Two additional Specific Aims were added to the Cooperative Agreement: **Specific Aim 4**, sequencing of *P. yoelii* to 3X coverage; **Specific Aim 5**, sequencing of *P. vivax* to 3X coverage. To date, we have published the first complete sequence of a malarial chromosome (*P. falciparum* chromosome 2), have sequenced and have nearly completed the sequences of *P. falciparum* chromosomes 10, 11, and 14. A plan for joiint annotation and publication of the *P. falciparum* genome with the Sanger Centre and Stanford University was prepared and implemented. Publication of the P. falciparum genome sequence is anticipated in late 2002. In addition, we completed sequencing of the rodent malaria *P. yoelii* to 5X coverage and have begun the sequencing of *P. vivax* to 5X coverage. Discussions with the Malaria Program, NMRC aimed at development of a program to use genomics and functional genomics to accelerate vaccine research are in progress.

References

1. 2001. PlasmoDB: An integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. The *Plasmodium* Genome Database Collaborative. Nucleic Acids Res **29:**66-9.
2. **Bloom, B. R.** 1995. A microbial minimalist. Nature **378:**236.
3. **Butler, D., J. Maurice, and C. O'Brien.** 1997. Briefing malaria. Nature **386:**535-540.
4. **Gardner, M. J.** 2001. A status report on the sequencing and annotation of the P. falciparum genome. Mol Biochem Parasitol **118:**133-8.
5. **Gardner, M. J., H. Tettelin, D. J. Carucci, L. M. Cummings, L. Aravind, E. V. Koonin, S. Shallom, T. Mason, K. Yu, C. Fujii, J. Pedersen, K. Shen, J. Jing, D. C. Schwartz, M. Pertea, S. Salzberg, L. Zhou, G. G. Sutton, R. L. Clayton, O. White, H. O. Smith, C. M. Fraser, M. D. Adams, J. C. Venter, and S. L. Hoffman.** 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science **282:**1126-1132.
6. **Jing, J., C. Aston, L. Zhongwu, D. J. Carucci, M. J. Gardner, J. C. Venter, and D. C. Schwartz.** 1999. Optical mapping of *Plasmodium falciparum* chromosome 2. Genome Research **9:**175-181.
7. **Kappe, S. H., M. J. Gardner, S. M. Brown, J. Ross, K. Matuschewski, J. M. Ribeiro, J. H. Adams, J. Quackenbush, J. Cho, D. J. Carucci, S. L. Hoffman, and V. Nussenzweig.** 2001. Exploring the transcriptome of the malaria sporozoite stage. Proc Natl Acad Sci U S A **98:**9895-900.
8. **Lai, Z., J. Jing, C. Aston, V. Clarke, J. Apodaca, E. T. Dimlanta, D. J. Carucci, M. J. Gardner, B. Mishra, T. Anantharaman, S. Paxia, S. L. Hoffman, J. C. Venter, E. J. Huff, and D. C. Schwartz.** 1999. A shotgun optical map of the entire *Plasmodium falciparum* genome. Nature Genetics **23:**309-313.
9. **Salzberg, S. L., M. Pertea, A. Delcher, M. J. Gardner, and H. Tettelin.** 1999. Interpolated Markov models for eukaryotic gene finding. Genomics **59:**24-31.

Appendices

Appendix A. Reprint: **Gardner, M. J.** 2001. A status report on the sequencing and annotation of the P. falciparum genome. Mol Biochem Parasitol **118:**133-8.

Appendix B. Reprint: **Kappe, S. H., M. J. Gardner, S. M. Brown, J. Ross, K. Matuschewski, J. M. Ribeiro, J. H. Adams, J. Quackenbush, J. Cho, D. J. Carucci, S. L. Hoffman, and V. Nussenzweig.** 2001. Exploring the transcriptome of the malaria sporozoite stage. Proc Natl Acad Sci U S A **98:**9895-900.

Appendix A

ELSEVIER

Molecular & Biochemical Parasitology 118 (2001) 133–138

MOLECULAR & BIOCHEMICAL PARASITOLOGY

www.parasitology-online.com.

Review

A status report on the sequencing and annotation of the *P. falciparum* genome

Malcolm J. Gardner *

The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Abstract

Almost 5 years ago, an international consortium of sequencing centers and funding agencies was formed to sequence the genome of the human malaria parasite *Plasmodium falciparum*. A novel chromosome by chromosome shotgun strategy was devised to sequence this very AT-rich genome. Two of the 14 chromosomes have been completed and the remaining chromosomes are in the final stages of gap closure. The consortium recently developed plans for the annotation and analysis of the complete genome sequence and its publication in 2002. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Genome; *Plasmodium falciparum*; Chromosome; Malaria

1. Introduction

The first complete genome sequence of a free-living organism, *Haemophilus in uenzae* , was published in 1995 [1]. Besides proving the speed and cost-effectiveness of the whole genome shotgun (WGS) approach to genome sequencing, this work introduced many scientists to the value of a complete genome sequence in terms of providing insights into the biology, biochemistry, and pathogenicity of microorganisms that cause disease. Several other genome sequences were completed soon after, and today, at least 55 microbial genomes have been sequenced, including both pathogenic and non-pathogenic organisms. As once predicted [2], only a few years after the completion of the first microbial genome sequence, scientists working on many of the most important human pathogens have entered the 'post-genomic era of microbe biology', and are building upon the foundation provided by complete genome sequences to drive research into the development of new drugs and vaccines against these organisms.

Within a few months of the publication of the *H. in uenzae* genome, several groups working on the human malaria parasite *P. falciparum* began to investigate the feasibility of determining its genome sequence. At the time, this seemed a daunting and perhaps impossible task. At an estimated 30 megabases (Mb), the *P. falciparum* genome was thought to be approximately 15-fold larger than that of *H. in uenzae* , so large that the ~500,000 shotgun sequences required could not have been assembled with the existing assembly software (the genome is now known to be about 25 Mb [3]). In addition, the genome is very AT-rich, and most investigators working on *P. falciparum* were all too familiar with the difficulty of cloning the DNA in *E. coli*, where it was frequently subject to deletions and rearrangements that precluded construction of high-quality, large insert genomic libraries. Were these deletions and rearrangements to occur in the libraries used for sequencing, it would have been impossible to obtain the complete genome sequence. Large fragments (> 100 kb) of *P. falciparum* DNA had been cloned in yeast artificial chromosomes and were relatively stable [4,5], but it was not possible to subclone short fragments of the YAC inserts without a great deal of cross-contamination with yeast DNA, making the YAC libraries unsuitable for large scale sequencing. Another major

* Tel.: +1-301-838-3519; fax: +1-301-838-0208.
E-mail address: gardner@tigr.org (M.J. Gardner).

PII: S0166-6851(01)00390-5

concern was the projected price of the project. With the existing techniques and costs, it was estimated that sequencing of *P. falciparum* would require at least $15 million, a sum not easily obtained from the usual funding sources.

Further discussions amongst members of the malaria research community, the sequencing centers, and representatives from the Wellcome Trust, the National Institute for Allergy and Infectious Diseases, the Burroughs Wellcome Fund, and the U.S. Department of Defense culminated in the formation of an international consortium to sequence the genome of *P. falciparum* [6]. Sequencing was to be conducted by the Pathogen Sequencing Unit at the Sanger Center, the Stanford University Genome Sequencing Center, and The Institute for Genomic Research (TIGR) and the Malaria Program at the Naval Medical Research Center (NMRC). Start-up funds were obtained for projects to investigate various sequencing strategies and to develop reagents prior to initiation of a full-scale effort. Ultimately, a chromosome by chromosome shotgun strategy was devised whereby the 14 chromosomes were purified on pulsed field gels and sequenced individually using a shotgun approach similar to that used for bacterial genomes. This strategy enabled the genome to be divided among the three sequencing centers and partitioned the genome into more manageable segments for assembly and gap closure. The consortium also organized a series of semi-annual meetings beginning in December 1996 [6]. These meetings provided a forum for the sharing of technical information, review and coordination of sequencing and related activities, and development of a data use policy for the use of preliminary sequence data released by the sequencing centers. These meetings have continued to this day, but as the *P. falciparum* sequencing effort gained momentum the meetings evolved to cover such topics as genome databases, functional genomics, and comparative genomics of apicomplexans.

2. Strategy and methodology

The chromosome by chromosome shotgun strategy proved to be fairly effective in the sequencing of *P. falciparum*, although the extreme AT-richness of the genome made the closure process extremely difficult. Briefly, the chromosomes of *P. falciparum* clone 3D7 were resolved on pulsed field gels and chromosomal DNA was extracted by agarase digestion. The DNA was then sheared into 1–2 kb fragments, cloned into plasmid or M13 vectors, and randomly-picked clones were sequenced. Chromosomes 2, 10, 11 and 14 were assigned to TIGR and the NMRC, chromosome 12 to the Stanford group, and the remaining chromosomes were assigned to the Sanger Centre, including the 'blob' of mid-sized chromosomes that could not be resolved on gels. Most of the sequence reactions were performed on ABI 377 slab gel sequencers using dye-terminator chemistry. The sequences were assembled to form contigs using either phrap (www.phrap.org) or TIGR Assembler [7]. The Sanger and Stanford groups also performed low pass sequencing of shotgun libraries prepared from YAC clones previously localized on the chromosomes by the Wellcome Trust Malaria Genome Collaboration [8]. The YAC-derived sequences were used to 'bin' the sequences obtained from the chromosomal libraries into smaller subsets prior to assembly. After assembly, contigs from adjoining regions of the chromosomes were identified by means of forward–reverse links [1], and the groups of linked contigs were mapped to the chromosomes by means of STSs [8,9] or microsatellite markers [10,11], and by comparison to optical restriction maps of the chromosomes [3,12]. Most of the techniques used to close the remaining gaps were basically the same used for other genome projects [1], such as primer walking along plasmid templates that crossed sequence gaps, and closure of physical gaps by PCR amplification and sequencing of genomic DNA fragments that spanned the gaps. Other techniques, however, had to be devised to assist in the closure of very AT-rich regions. Many parts of the genome, such as the putative centromeric sequences identified on chromosomes 2 and 3 by Bowman et al. [13], were over 97% AT, and many regions in the vicinity of long runs of A's and T's proved very difficult to sequence accurately. In these cases, transposon insertion [14] or microlibrary techniques were used to generate a high sequence coverage across the AT-rich area, from which a more accurate sequence could be obtained (potential secondary structures in AT-rich areas that may have interfered with sequencing may also have been disrupted by insertion of a transposon or shotgunning of a fragment during microlibrary construction). These procedures are very labor intensive and time-consuming, however, and dealing with these AT-rich areas is one reason why closure of the *P. falciparum* genome has taken such a long time. Once whole chromosomes or substantial contigs were completed, the sequences were edited to resolve any ambiguities. Optical restriction maps [3,12] proved invaluable for verification that the chromosome sequences had been assembled correctly [14].

3. First glimpses of the *P. falciparum* genome

Completion of the first two chromosome sequences provided detailed pictures of chromosome organization and fascinating previews of the *P. falciparum* genome [14,13]. Some of the major findings included the discovery of two new gene families that were predicted to

encode potentially variant surface antigens (rifins and STEVORS [14–17]), a cluster of four genes of unknown function repeated on one end of chromosomes 2 and 3 [13], genes encoding enzymes of the type II fatty acid biosynthetic pathway previously thought to be restricted to plants and bacteria [14,18], and putative centromeres [13]. Gene density was just under 1 gene per 5 kb, very similar to *C. elegans*, and approximately one-half of genes were predicted to contain introns, although recent studies indicate this may have been an underestimate. Almost two-thirds of the predicted genes had no detectable orthologs in other organisms, suggesting that many aspects of parasite biology have not yet been uncovered despite many years of research.

In addition to the data release associated with the publication of chromosomes 2 and 3, preliminary contigs for all chromosomes have been released periodically, virtually since the beginning of the project, and have been accessible at the sequencing centers' web sites, at NCBI, and more recently at a new community database for malaria genome information, PlasmoDB [19] (www.plasmodb.org). Preliminary annotation is also available at these sites. A data release policy devised by the sequencing centers and the funding agencies, with input from members of the malaria research community, was also established. Although somewhat controversial [20–22], the data release policy allowed many scientists around the world to get early glimpses of the genome sequence data and 'provide ... information that may jump-start biological experimentation' (www.tigr.org/tdb/edb2/pfa1/htmls/), while protecting the right of the sequencing centers to publish whole chromosome or whole genome analyses of the data they had so laboriously produced. Dozens of reports have since been published in which use of the preliminary sequence data was acknowledged. Virtually every area of malaria biology and biochemistry has been positively affected by the release of preliminary genome sequence information. Some of the most outstanding discoveries have been the identification of new drug targets, opening new avenues for the development of novel antimalarials [23–26]. Many other research projects that rely in some fashion on the preliminary sequence data are underway, including the development of full-genome microarrays and proteomics studies.

4. Current status and plans for annotation

The consortium met at the Sanger Centre in June 2001 to review progress in gap closure and make plans for annotation and publication of the *P. falciparum* genome sequence. Chromosomes 2 and 3 have been published. Chromosomes 1, 4, 9–12 and 14 were reported to be in the final stages of closure, with only a handful of gaps per chromosome remaining. Chromosomes in the 'blob' (chromosomes 5–8) and chromosome 13, have full sequence coverage but are lagging behind in gap closure, and still consist of hundreds of contigs per chromosome. Gap closure has been slow due to the paucity of markers for ordering of the contigs. However, the Sanger Centre recently began Happy Mapping [27,28] of the contigs and should complete this task in the fall of 2001. Once these contigs have been grouped and ordered along the chromosomes, gap closure for these chromosomes is expected to accelerate.

The sequencing centers also laid plans for annotation of the genome sequence, leading early in 2002, to submission of joint publication on the analysis of the entire *P. falciparum* genome and a series of papers on the chromosomes by the three sequencing groups. The basic elements of this plan include beginning the annotation on a set of contig sequences representing the best available data for each chromosome. These contigs would be 'frozen' so to permit annotation to proceed on a stable data set, and where possible the contigs will be joined end-to-end in the correct order and orientation to form draft chromosome sequences. As the annotation of these draft chromosomes is underway, closure efforts on the remaining gaps will continue, and the new sequence data generated during the closure process will be merged into the annotated contigs near the end of the process. Each sequencing center will be responsible for annotation of the chromosomes they sequenced, using the software and methods in use at each center. In an attempt to ensure that the annotation done by the participating centers is of equal quality, the same 100 kb sequence will be annotated by the three groups early in the annotation process and the results will be compared to identify any problems. Furthermore, it was agreed that TIGR will maintain a central relational database containing a representation of the sequence data and annotation produced at all three centers, and that the centers will develop procedures for the frequent semi-automated exchanges of data. This will allow all of the annotators to view the same picture of the complete genome and facilitate whole genome analyses. Importantly, this arrangement will also simplify the process of submitting the annotated genome sequence to the PlasmoDB database [19]. This plan has now been put in motion. Many chromosome sequences have been frozen, annotation has begun, and the system for data exchange between centers is being tested.

As with annotation of chromosomes 2 and 3, and other eukaryotic genomes, including the human genome [29–31], annotation of the complete *P. falciparum* genome presents many challenges. A major problem is the difficulty of gene prediction in eukaryotic genomes. Two gene finders specifically designed to predict gene models in the gene-dense *P. falciparum* genome are now available, GlimmerM [32] and phat [33]. Both programs

perform well but predict different gene models in some cases. The human annotator, faced with conflicting models and in many instances with no other evidence such as EST hits or protein matches to confirm either model, has great difficulty in deciding which model, if any, is likely to be the correct one. Subjective criteria (otherwise known as 'the force') must sometimes be employed in selecting one model over another. Another problem is that the gene finders do not detect genes in some regions of the genome where they would be expected to occur, suggesting that some genes may have escaped detection. It was for this reason that the systematic gene nomenclature system devised for *P. falciparum* numbers the genes in increments of five, to allow genes identified later to be neatly inserted into the annotation [14]. The gene finders are also unable to handle the complexities of alternative splicing. In short, gene models are predictions, and investigators using the annotation would be wise to verify the gene models experimentally prior to embarking on detailed studies of these genes. In an attempt to improve gene modeling during the whole genome annotation that is just beginning, the original training set used for GlimmerM [32] was recently updated to include experimentally-verified genes published over the past 3 years, and both GlimmerM and phat have been re-trained on new training set. EST datasets from a variety of organisms have also been updated [34] (www.tigr.org/tdb/tgi.html); these can provide the annotator with experimental evidence to support complete gene models or intron predictions. Genome annotation is an ongoing process, and the upcoming annotation of the complete *P. falciparum* genome should be viewed as the first step in a process that will continue for many years. Continual feedback from the malaria research community, in the form of experimentally verified gene structures, will be essential in order to improve the genome annotation process.

One major improvement over the previous annotation of chromosomes 2 and 3 will be in the assignment of genes into functional role categories. For chromosomes 2 and 3, existing role categories that had been devised for prokaryotes were adapted for use with a eukaryotic organism. Both centers used different schemes, and neither scheme captured the increased complexity of eukaryotic biology. To avoid this situation, the whole genome annotation will use the Gene Ontology (GO) system that is currently used by several organism-specific databases including the *Saccharomyces cerevisiae* database SDB and FlyBase, among others [35]. The GO system consists of three separate ontologies (molecular function, biological process, and cellular component), each with 'a set of structured vocabularies ... that can be used to describe gene products in any organism [36]. A group of parasitologists, coordinated by Matt Berriman at the Sanger Centre, is currently drafting a set of defined terms that describe novel aspects of parasite biology for inclusion into the GO system (e.g. the term 'rosetting' in the biological process ontology). Thus, the *P. falciparum* annotation will use a more powerful and widely utilized gene system of gene product classification, enabling users to gain broader insights into parasite biology from the annotated genome sequence.

5. Sequencing of additional *P. falciparum* clones and *Plasmodium* spp.

The success of the *P. falciparum* sequencing effort led many investigators to call for sequencing of additional *Plasmodium* spp. and a more recent isolate(s) of *P. falciparum*. Of particular interest was generation of sequence data for many of the malaria parasites used as model systems for drug and vaccine development, and for *Plasmodium vivax*, the second most important human malaria parasite. Although the chromosome by chromosome approach to sequencing of *P. falciparum* has been successful, there are a number of reasons why it would be best to avoid this strategy when additional *Plasmodium* spp. are sequenced. One, the introduction of capillary-based sequencers such as the ABI 3700 has dramatically increased the sequencing capacity of genome centers while simultaneously lowering the costs of sequencing. At the time the malaria genome project was started, sequencing of a 30 Mb genome would have been a very large project. Today, the random sequences required for such a project can be generated much more quickly and at lower expense than even 2 years ago, so that dividing the sequencing of a 30 Mb genome between several centers would not be required for purely logistical reasons. In addition, a major problem with the slab gel sequencers was the high frequency of mistracked sequences, which interfered with the assembly and contig grouping procedures. Mistracking does not occur with the capillary based sequencers in which each reaction is contained within a single capillary during electrophoresis. Two, preparations of pulsed field gel purified chromosomal DNA were always cross-contaminated with DNA from other chromosomes. For chromosome 2, we estimated that 20% of the sequences obtained were from other regions of the genome. The cross-contamination resulted in the formation of many short, low coverage contigs that confounded the gap closure process, and since every separate chromosome project generated its own set of 'contaminants,' more sequences had to be generated in the chromosome by chromosome approach to produce the required sequence coverage of a chromosome than might have been required using the whole genome strategy. Third, the chromosome by chromosome strategy was necessitated, in part, by the inability of the existing assembly software to assemble the $\sim$500,000 shotgun sequences

that would have been produced by a whole genome shotgun approach. In fact, the version of the TIGR Assembler that was available early in the chromosome 2 pilot project had difficulty assembling 9000 sequences in a reasonable time frame. More recent versions of TIGR Assembler and new assemblers such as the Celera Assembler [37] can handle much larger data sets. In summary, future efforts to sequence another clone of *P. falciparum*, perhaps from a recent clinical isolate, or another species of *Plasmodium*, could be done using a whole genome approach at any one of several sequencing centers. Several such efforts are already underway or in the planning stages, including the sequencing of *P. yoelii* and *P. vivax* to 5× coverage (TIGR/NMRC), and five other *Plasmodium* spp. to 3× coverage (*P. chabaudi*, *P. berghei*, *P. knowlesi*, *P. reichenowi*, and *P. gallinaceum*) by the Sanger Centre. These projects should produce contigs of 2–5 kb representing >90% of the parasites' genomes. Besides providing gene sequences that can be used to facilitate a variety of functional studies, these projects will allow comparative genome analyses of *Plasmodium* spp. that have very different biological characteristics. Several other apicomplexan parasites are also being sequenced, including *Theileria parva* (TIGR and the International Livestock Research Institute), *T. annulata* (The Sanger Centre), and two isolates of *Cryptosporidium parvum* (University of Minnesota and the Medical College of Virginia).

6. Summary

After 5 years of extraordinary effort by the consortium, completion of the *P. falciparum* genome sequence appears imminent. Analysis of the first two chromosomes to be sequenced, which together represented about 8% of the genome, and exciting findings that were made possible by release of preliminary sequence data, have already justified the efforts made to sequence the genome of this deadly parasite. Annotation of the genome sequence has begun following a plan devised be the three sequencing centers and publication of an analysis of the *P. falciparum* genome is expected in 2002. The success of the *P. falciparum* project has spawned similar efforts to determine the genome sequences of additional *Plasmodium* spp. and other apicomplexans. In addition, the human genome sequence [29,30], and the *Anopheles gambiae* genome sequence that is also expected to be completed in 2002 (www.niaid.nih.gov/newsroom/releases/celera.htm), provide opportunities for study of host–vector–parasite relationships. In the years to come, the complete genome sequences of all three members of the *Plasmodium* life cycle will allow investigators to gain a better understanding of parasite biology and will be invaluable resources in the quest to develop new drugs and vaccines to fight malaria.

Acknowledgements

I thank my colleagues at TIGR, the Naval Medical Research Center, and the members of the Malaria Genome Sequencing Consortium at the Sanger Centre and the Stanford Genome Technology Center, for their support. Sequencing of the *P. falciparum* genome at TIGR and the NMRC is supported by the National Institute of Allergy and Infectious Diseases (U01 AI42243), the Burroughs Wellcome Fund (990785), the Department of the Army (DAMD17-98-2-8005), and Naval Medical Research Center Work Unit Nos. 61102A.S13.00101.BFX1431, 612787A.870.00101.EFX.1432, 623002A.810.00101.HFX.1433 and STEP C611102A0101BCX.

References

[1] Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus in uenzae* Rd. Science 1995;269:496–512.
[2] Bloom BR. A microbial minimalist. Nature 1995;378:236.
[3] Lai Z, Jing J, Aston C, et al. A shotgun optical map of the entire *Plasmodium falciparum* genome. Nat Genet 1999;23:309–13.
[4] Triglia T, Kemp DJ. Large fragments of *Plasmodium falciparum* DNA can be stable when cloned in yeast artificial chromosomes. Mol Biochem Parasitol 1991;44:207–11.
[5] de Bruin D, Lanzer M, Ravetch JV. Characterization of yeast artificial chromosomes from *Plasmodium falciparum*: construction of a stable, representative library and cloning of telomeric DNA fragments. Genomics 1992;14:332–9.
[6] Hoffman SL, Bancroft WH, Gottlieb M, et al. Funding for malaria genome sequencing. Nature 1997;387:647.
[7] Sutton GS, White O, Adams MD, et al. TIGR assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci Technol 1995;1:9–19.
[8] Foster J, Thompson J. The *Plasmodium falciparum* genome project: a resource for researchers. Parasitol Today 1995;11:1–4.
[9] Dame JB, Arnot DE, Bourke PF, et al. Current status of the *Plasmodium falciparum* genome project. Mol Biochem Parasitol 1996;79:1–12.
[10] Su XZ, Wellems TE. Toward a high-resolution *Plasmodium falciparum* linkage map:polymorphic markers from hundreds of simple sequence repeats. Genomics 1996;33:430–44.
[11] Su XZ, Wellems TE. *Plasmodium falciparum*: assignment of microsatellite markers to chromosomes by PFG-PCR. Exp Parasitol 1999;91:367–9.
[12] Jing J, Aston C, Zhongwu L, et al. Optical mapping of *Plasmodium falciparum* chromosome 2. Genome Res 1999;9:175–81.
[13] Bowman S, Lawson D, Basham D, et al. The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature 1999;400:532–8.
[14] Gardner MJ, Tettelin H, Carucci DJ, et al. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science 1998;282:1126–32.
[15] Cheng Q, Cloonan N, Fischer K, et al. Stevor and rif are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. Mol Biochem Parasitol 1998;97:161–76.
[16] Fernandez V, Hommel M, Chen Q, et al. Small, clonally variant antigens expressed on the surface of the *Plasmodium falciparum*-infected erythrocyte are encoded by the rif gene family and are

the target of human immune responses. J Exp Med 1999;190:1393–404.
[17] Kyes SA, Rowe JA, Kriek N, et al. Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. Proc Natl Acad Sci USA 1999;96:9333–8.
[18] Waller RF, Keeling PJ, Donald RGK, et al. Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. Proc Natl Acad Sci USA 1998;95:12 352–2 357.
[19] The Plasmodium Genome Database Collaborative. PlasmoDB: an integrative database of the *Plasmodium falciparum* genome. Tools for accessing and analyzing finished and unfinished sequence data. Nucleic Acids Res 2001;29:66–9.
[20] Gottlieb M, McGovern V, Goodwin P, et al. Please don't downgrade the sequencers' role. Nature 2000;406:121–2.
[21] Macilwain C. Biologists challenge sequencers on parasite genome publication. Nature 2000;405:601–2.
[22] Pace T. When public-interest science needs solidarity. Nature 2000;406:122.
[23] Jomaa H, Wiesner J, Sanderbrand S, et al. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science 1999;285:1573–6.
[24] Vollmer M, Thomsen N, Wiek S, et al. Apicomplexan parasites possess distinct nuclear-encoded, but apicoplast-localized, plant-type ferredoxin-NADP + reductase and ferredoxin. J Biol Chem 2001;276:5483–90.
[25] Lee CS, Salcedo E, Wang Q, et al. Characterization of three genes encoding enzymes of the folate biosynthetic pathway in *Plasmodium falciparum*. Parasitology 2001;122(Pt. 1):1–13.
[26] Salcedo E, Cortese JF, Plowe CV, et al. A bifunctional dihydrofolate synthetase—folylpolyglutamate synthetase in *Plasmodium falciparum* identified by functional complementation in yeast and bacteria. Mol Biochem Parasitol 2001;112:239–52.
[27] Dear PH, Cook PR. Happy mapping: linkage mapping using a physical analogue of meiosis. Nucleic Acids Res 1993;21:13–20.
[28] Piper MB, Bankier AT, Dear PH. A HAPPY map of *Cryptosporidium parvum*. Genome Res 1998;8:1299–307.
[29] Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304–51.
[30] Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
[31] Hogenesch JB, Ching KA, Batalov S, et al. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell 2001;106:413–5.
[32] Salzberg SL, Pertea M, Delcher A, et al. Interpolated Markov models for eukaryotic gene finding. Genomics 1999;59:24–31.
[33] Cawley SE, Wirth AI, Speed TP. Phat—a gene finding program for *Plasmodium falciparum*, Mol Biochem Parasitol, in press.
[34] Carlton JM-R, Muller R, Yowell CA, et al. Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species, Mol Biochem Parasitol 2001:118;201–210.
[35] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–9.
[36] The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. Genome Res 2001;11:1425–33.
[37] Huson DH, Reinert K, Kravitz SA, et al. Design of a compartmentalized shotgun assembler for the human genome. Bioinformatics 2001;17(Suppl. 1):S132–9.

Exploring the transcriptome of the malaria sporozoite stage

Stefan H. I. Kappe*†, Malcolm J. Gardner‡, Stuart M. Brown§, Jessica Ross*, Kai Matuschewski*, Jose M. Ribeiro¶, John H. Adams‖, John Quackenbush‡, Jennifer Cho‡, Daniel J. Carucci**, Stephen L. Hoffman††, and Victor Nussenzweig*

*Michael Heidelberger Division, Department of Pathology, Kaplan Cancer Center, New York University School of Medicine, New York, NY 10016; ‡The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850; §Research Computing Resource, New York University Medical Center, New York, NY 10016; ¶Medical Entomology Section, Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892-0425; ‖Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556; **Malaria Program, Naval Medical Research Center, Silver Spring, MD 20910; and ††Celera Genomics, 45 West Gude Drive, Rockville, MD 20850

Edited by Louis H. Miller, National Institutes of Health, Bethesda, MD, and approved June 19, 2001 (received for review April 13, 2001)

Most studies of gene expression in *Plasmodium* have been concerned with asexual and/or sexual erythrocytic stages. Identification and cloning of genes expressed in the preerythrocytic stages lag far behind. We have constructed a high quality cDNA library of the *Plasmodium* sporozoite stage by using the rodent malaria parasite *P. yoelii*, an important model for malaria vaccine development. The technical obstacles associated with limited amounts of RNA material were overcome by PCR-amplifying the transcriptome before cloning. Contamination with mosquito RNA was negligible. Generation of 1,972 expressed sequence tags (EST) resulted in a total of 1,547 unique sequences, allowing insight into sporozoite gene expression. The circumsporozoite protein (CS) and the sporozoite surface protein 2 (SSP2) are well represented in the data set. A BLASTX search with all tags of the nonredundant protein database gave only 161 unique significant matches ($P(N)$ 10^{-4}), whereas 1,386 of the unique sequences represented novel sporozoite-expressed genes. We identified ESTs for three proteins that may be involved in host cell invasion and documented their expression in sporozoites. These data should facilitate our understanding of the preerythrocytic *Plasmodium* life cycle stages and the development of preerythrocytic vaccines.

Plasmodium yoelii yoelii | expressed sequence tag

Protozoan parasites of the genus *Plasmodium* are the causative agents of malaria, the most devastating parasitic disease in humans. The parasites occur in distinct morphological and antigenic stages as they progress through a complex life cycle, thwarting decades of efforts to develop an effective malaria vaccine. *Plasmodium* is transmitted via the bite of an infected *Anopheles* mosquito, which releases the sporozoite stage into the skin. Sporozoites enter the bloodstream and, on reaching the liver, invade hepatocytes and develop into exo-erythrocytic forms (EEF). After multiple cycles of DNA replication, the EEF contains thousands of merozoites (liver schizont) that are released into the blood stream and initiate the erythrocytic cycle (asexual blood stage) that causes the disease malaria. Changes in life cycle stages are accompanied by major changes in gene expression and therefore by major changes in antigenic composition. The form of the parasite best studied is the asexual blood stage, mainly because of its comparatively easy experimental accessibility. Therefore, most *Plasmodium* proteins that have been well characterized are expressed during the erythrocytic cycle, among them some major erythrocytic-stage vaccine candidates such as merozoite surface protein-1 (MSP-1) and apical membrane antigen-1 (AMA-1; ref. 1). Erythrocytic-stage vaccines are aimed at inducing an immune response that suppresses or eradicates parasite load in the blood. In contrast, preerythrocytic vaccines are aimed at eliciting an immune response that destroys the sporozoites and the EEF, thereby preventing progression of the parasite to the blood stage. The feasibility of a preerythrocytic vaccine is demonstrated by the fact that immunization with radiation-attenuated sporozoites leads to protective, sterile immunity (2, 3). The effector mechanisms are antibodies (4), cytotoxic T lymphocytes (CTL; ref. 4), and lymphokines (5, 6). Hence, it is desirable to systematically identify proteins synthesized by sporozoites and EEF to select new potential vaccine candidates. Antibodies against surface-exposed sporozoite proteins block hepatocyte entry (7). In addition, sporozoite proteins can be carried over into the invaded hepatocyte and become a target for CTL (8). By using mixtures of these proteins, it might be possible to formulate a vaccine that mimics the sterile immunity achieved by immunization with irradiated sporozoites. Sporozoite proteins could also be the target of transmission-blocking strategies. Past efforts to prepare cDNA libraries of sporozoites and identify new sporozoite antigens were hindered by difficulties in obtaining adequate numbers of purified parasites. Thus far, few sporozoite-expressed proteins have been identified. The best characterized of these proteins are the circumsporozoite protein (CS; ref. 2) and the sporozoite surface protein 2 (SSP2), also called thrombospondin-related anonymous protein (TRAP; refs. 9–11). CS and SSP2/TRAP are involved in the invasion of hepatocytes and are detected in the hepatocyte after sporozoite invasion. Both proteins are found in all *Plasmodium* species examined. A few other sporozoite antigens have been identified in *P. falciparum* (12, 13), but their function is unknown.

To facilitate the identification of genes that are expressed in the sporozoite stage, we have constructed a cDNA library from salivary gland sporozoites of the rodent malaria parasite *Plasmodium yoelii* and generated 1,972 expressed sequence tags (ESTs). We document the quality of the library by the presence of CS and SSP2/TRAP transcripts and the absence of erythrocytic stage-specific transcripts. The sequence data provide insight into sporozoite gene expression. We show sporozoite expression of MAEBL (14), a protein previously thought to be present only in erythrocytic stages. In addition, we identify two putative sporozoite adhesion ligands. Transcripts of a key enzyme of the shikimate pathway (15) are present in the data set, indicating that this pathway is likely to be operational in sporozoites and liver stages.

MICROBIOLOGY

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: CS, circumsporozoite protein; SSP2, sporozoite surface protein 2; TRAP, thrombospondin-related anonymous protein; EST, expressed sequence tag; EEF exo-erythrocytic form; MSP-1, merozoite surface protein-1; MyoA, myosin A; TSR, thrombospondin type 1 repeat; SPATR, secreted protein with altered thrombospondin repeat.

Data deposition: The EST sequences reported in this paper have been deposited in the GenBank dbEST database (accession nos. BG601070–BG603042). Complete gene sequences have been deposited in the GenBank database (accession nos. AF390551–AF390553).

†To whom reprint requests should be addressed. E-mail: kappes01@popmail.med.nyu.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Materials and Methods

Parasite Preparation. Two million *P. yoelii* (17XNL) sporozoites were obtained in a salivary gland homogenate from dissection of 100 infected *Anopheles stephensi* mosquitos. The crude salivary gland homogenate was passed over a DEAE cellulose column to remove contaminating mosquito tissue. Sporozoites (4×10^5) were recovered after purification. The preparation was almost free of mosquito contaminants as judged by microscopic inspection. Sporozoites were immediately subjected to poly(A)$^+$ RNA extraction.

RNA Extraction and cDNA Synthesis. Poly(A)$^+$ RNA was directly isolated from the sporozoites by using the MicroFastTrack procedure (Invitrogen) and was resuspended in a final volume of 10 μl elution buffer (10 mM Tris, pH 7.5). The obtained poly(A)$^+$ RNA was treated with Dnase I (Life Technologies, Rockville, MD) to remove possible genomic DNA contamination. RNA quantification was not possible because of the minute amounts obtained. The RNA was reverse-transcribed by using Superscript II (Life Technologies), a modified oligo(dT) oligonucleotide for first strand priming (5′-AAGCAGTGGTAACAACGCAGAGTACT$_{30}$VN-3′; V = A/C/G, N = A/C/G/T) and a primer called cap switch oligonucleotide (5′-AAGCAGTGGTAACAACGCAGAGTACGCGGG-3′) that allows extension of the template at the 5′ end (CLONTECH). Second strand synthesis and subsequent PCR amplification was done with an oligonucleotide that anneals to both the modified oligo(dT) oligonucleotide and the cap switch oligonucleotide.

cDNA Cloning and Sequencing. The cDNA was size selected on a CHROMA-SPIN 400 column (CLONTECH) that resulted in a cutoff at ≥300 bp and was ligated into vector pCR4 (Invitrogen). Ligations were transformed into *Escherichia coli* TOP10-competent cells. Template preparation and sequencing were done as described (16). Sequencing was performed in both directions.

Assemblies and Database Searches. All obtained sequences were subjected to vector sequence removal and screened for overlaps, and matching sequences were then assembled by using the TIGR assembler program. The nonredundant (NR) sequence database at the National Center for Biotechnology Information (NCBI) was searched with the complete data set, consisting of the assembled sequences and singletons, by using the Basic Local Alignment Search Tool X (BLASTX) algorithm.

Sources of Sequence Data. Sequence data were obtained from the TIGR *P. yoelii* genome project (www.tigr.org) and the *Plasmodium* genome consortium PlasmoDB (http://PlasmoDB.org).

cDNA Blots. cDNA was separated on agarose gels and transferred to nylon membranes (Roche). Gene-specific probes were prepared by using the digoxigenin (DIG) High Prime Labeling system (Roche). cDNA blots were incubated and washed according to the manufacturer's instructions (Roche).

Reverse Transcription–PCR. Poly(A)$^+$ RNA was reverse-transcribed by using Superscript II. Gene-specific PCR was done by using oligonucleotide primers specific for *P. yoelii MSP-1* (L22551; sense, 5′-GGTAAAAGCTGGCGTCATTGATCC-3′; antisense, 5′-GTCTAATTCAAAATCATCGGCAGG-3′) or *P. yoelii MAEBL* (AF031886; sense, 5′-ATGCTGCTCAATATCAGATTATTGC-3′; antisense, 5′-AACAATTTCATCAAAAGCAACTTCC-3′).

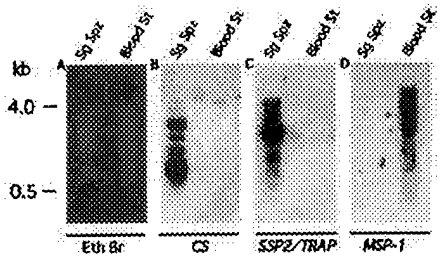

Fig. 1. Quality assessment of the generated cDNA populations. cDNA blot hybridization with stage-specific probes demonstrates that stage-specific transcript representation is not altered by cDNA amplification. (*A*) Ethidium bromide-stained agarose gel of cDNA amplified from salivary gland sporozoites (Sg Spz) or mixed blood stages (Blood St). Note the distinct bands visible in the sporozoite preparation. (*B*) Hybridization to a *CS* probe. (*C*) Hybridization to an *SSP2/TRAP* probe. (*D*) Hybridization to an *MSP-1* probe. Sizes are given in kb.

Indirect Immunofluorescence Assay. Salivary gland sporozoites and midgut sporozoites were incubated in 3% BSA/RPMI medium 1640 on BSA-covered glass-slides for 30 min, fixed, and permeabilized with 0.05% saponin. MAEBL was detected with the polyclonal antisera against the M2 domain or the 3′-carboxyl cysteine-rich region (1:200; ref. 14) and FITC-conjugated goat anti-rabbit IgG (1:100; Kirkegaard & Perry Laboratories).

Results

Quality Assessment of the cDNA Library. The amplified sporozoite cDNAs showed a visible size distribution between 300 and 4,000 bp on ethidium bromide-stained agarose gels, with highest density between 500 and 3,000 bp (Fig. 1*A*). No amplification was detected when the reverse transcription step was omitted (data not shown). To assess the quality of the sporozoite cDNA population, we performed cDNA blot analysis with probes for the sporozoite-expressed *SSP2/TRAP* and *CS*. cDNAs for both proteins were found to be abundant in salivary gland sporozoite preparations but absent in blood stage parasite preparations (Fig. 1 *B* and *C*). Conversely, cDNAs for the blood stage-expressed *MSP-1* were detected in blood stage parasite preparations but absent in sporozoites (Fig. 1*D*). The cDNA blot analysis documented the presence of cDNAs of the approximate full-length size of each transcript. In addition, smaller sized cDNA fragments were present for each transcript, resulting in multiple signals from distinctly sized cDNAs (Fig. 1). To assure that no trace amounts of genomic DNA were amplified, we analyzed the sporozoite cDNA for the presence of introns by using the transcript of myosin A (MyoA), a myosin that is expressed in the sporozoite stage (17). *MyoA* contains two introns, and neither was detected in the sporozoite cDNA preparation (data not shown). Sequencing of 100 clones confirmed the cDNA fragmentation, which was mainly due to internal priming by the modified oligo(dT) oligonucleotide. It annealed to homo-polymeric runs of adenine in the untranslated regions (UTR) and the coding sequences of this AT-rich organism. We took advantage of the AT-richness of the *P. yoelii* genome to differentiate between cDNAs of parasite origin and cDNAs amplified from contaminating mosquito RNA. Based on the total number of cDNA clones of mosquito origin, contamination was estimated to be ≈1%.

Characteristics of the EST Data Set. We obtained a final number of 1,972 sequence reads of sufficient quality to be subjected to further analysis (Table 1). The average length of EST sequence was 377 bp. Six hundred forty-eight of the sequence reads could be assembled into 223 consensus sequences (input files), and

Table 1. General characteristics of the *P. yoelii* sporozoite EST project

ESTs submitted to NCBI	1,972
ESTs in input files	648
Input files	223
Singletons	1,324
Total number of unique sequences	1,547
BLASTX matches	286
Unique BLASTX matches	161
Matches with proteins of unknown function	75
BLASTX matches with *Plasmodium* proteins	70
ESTs for CS	33
ESTs for SSP2/TRAP	13
ESTs for MAEBL	10
ESTs for HSP-70	10

1,324 sequences did not match another sequence in the data set sufficiently to allow assembly (singletons). This analysis gave a total of 1,547 unique sequences. A BLASTN comparison between the 1,547 unique sequences and the incomplete *P. yoelii* genome (2× coverage) database resulted in 1,135 matches. A BLASTX search of the predicted proteins from *the P. falciparum* genome (translated ORFs of >100 bases) resulted in only 356 matches, with a smallest sum probability of $P(N) \leq 10^{-4}$. A BLASTX search of the NR sequence database at NCBI resulted in only 286 matches, with a smallest sum probability of $P(N) \leq 10^{-4}$. Of those, 70 were matches with known *Plasmodium* proteins. The matches were grouped in functional categories shown in Fig. 2 (see Table 2, which is published as supplemental data on the PNAS web site, www.pnas.org, for a complete list of all BLASTX matches). All ESTs have been deposited in the GenBank dbEST database (accession nos. BG601070–BG603042). In addition, data are made available through the *P. yoelii* gene index (http://www.tigr.org/tdb/pygi/).

Functional Groups of ESTs. Ribosomal proteins were not very abundant, with only 7 of the estimated 80 components of the ribosome represented. Only 4 ESTs gave matches with other proteins involved in translation. This low representation of proteins of the translation machinery contrasts with the relative abundance of ribosomal proteins found in EST sequencing projects for *Toxoplasma* tachyzoites (12% of all ESTs; refs. 18 and 19) and *Cryptosporidium* sporozoites (8% of all ESTs; ref. 20). However, a *P. falciparum* blood stage parasite EST project found that proteins involved in translation were also underrepresented (21). There were 18 ESTs in the transcription category, 7 matching a *P. falciparum* RNA recognition motif binding protein and two matching a human zinc finger protein potentially involved in transcription.

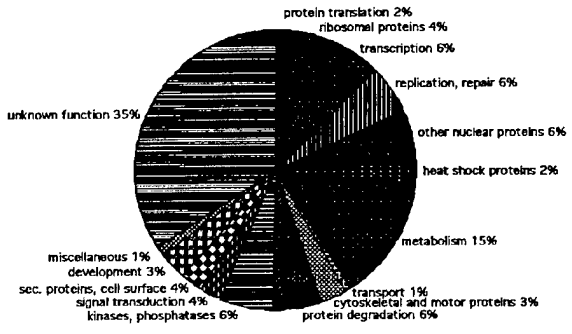

Fig. 2. Functional classification of *P. yoelii* sporozoite ESTs. One hundred sixty-one unique BLASTX matches were classified according to their putative biological function. Refer to Table 2 for a complete list of all BLASTX matches.

Especially significant among the ESTs giving BLASTX matches with proteins involved in metabolic pathways is chorismate synthase, the final enzyme of the shikimate pathway. This pathway generates the aromatic precursor chorismate, which is used for aromatic amino acid biosynthesis. The shikimate pathway is present in plants, fungi, and Apicomplexa (15) but is not found in vertebrates.

The salivary gland sporozoite is highly motile, and its main function is the invasion of the vertebrate hepatocyte. Of relevance to motility and invasion are tags for two apicomplexan unconventional class XIV myosins, MyoA and MyoB. MyoA localized under the plasma membrane within all invasive stages of *Plasmodium* (sporozoite, merozoite, and ookinete; refs. 17, 22, and 23), and a homologous protein was expressed in the *Toxoplasma* tachyzoite (24, 25). This myosin is currently the best candidate for the motor protein that drives Apicomplexan motility and host cell penetration.

Kinases and phosphatases are likely to be involved in the regulation of motility and host cell invasion (26), and we find 10 different input files and singletons in this category. Recently it was shown that a calmodulin-domain kinase, represented with one EST in the data set, played a crucial role in *Toxoplasma* tachyzoite motility and host cell invasion (27). Phospholipase A_2 is represented with one EST. Involvement of secreted phospholipase A_2 in the invasion process was shown in *Toxoplasma* tachyzoites (28). It will be of interest to find out whether this *Plasmodium* homologue has a role in hepatocyte invasion and/or plays a role in the migration of sporozoites through cells before establishing an infection (29).

The group of predicted secreted proteins and proteins that have a membrane anchor are of special interest, because they may be involved in host cell recognition and/or invasion. Within this group is the CS protein, most likely glycosylphasphatidylinositol-anchored, and SSP2/TRAP, a type one transmembrane protein. CS had one of the highest representations in the EST set with 33 matches, and TRAP was represented with 13 matches (Table 1).

Identification of Three Potential Sporozoite Invasion Ligands. Unexpectedly, we found that MAEBL was represented with 10 ESTs (Table 1). It was reported previously that MAEBL is expressed in *P. yoelii* and *P. berghei* merozoites, where it localized to the rhoptry organelles (14, 30). MAEBL is a type one transmembrane protein with a chimeric structure. It shares similarity with apical membrane antigen-1 (AMA-1) in the N-terminal portion, and similarity with the erythrocyte binding protein (EBP) family in the C-terminal portion (31). To ensure that the representation of a merozoite rhoptry protein in our EST library was not an artifact, we hybridized a salivary gland and midgut sporozoite cDNA blot to a *MAEBL*-specific probe, resulting in strong signals for both populations (Fig. 3*A*). In addition, reverse transcription–PCR with gene-specific primers resulted in *MAEBL* amplification from salivary gland sporozoite poly(A)$^+$ RNA and from blood stage poly(A)$^+$ RNA. In contrast, *MSP-1* expression was detected only in blood stages (Fig. 3*B*). A polyclonal antiserum against the carboxyl cysteine-rich region of *P. yoelii* MAEBL strongly reacted with permeabilized *P. yoelii* salivary gland sporozoites and midgut sporozoites in indirect immunofluorescence assay (IFA), indicating that this protein is indeed expressed in the sporozoite stages (Fig. 3 *C* and *D*). MAEBL localization was heterogeneous but was frequently more pronounced in one end of the sporozoites. Similar staining was obtained with a polyclonal antiserum against the M2 domain of MAEBL (data not shown).

MICROBIOLOGY

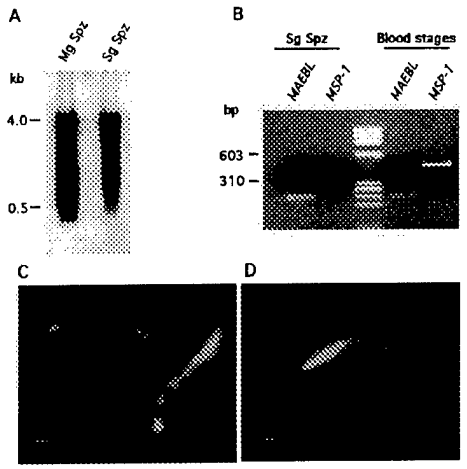

Fig. 3. Sporozoite expression of MAEBL. (*A*) cDNA blot showing *MAEBL* expression in midgut sporozoites (Mg Spz) and salivary gland sporozoites (Sg Spz). (*B*) Reverse transcription–PCR confirming *MAEBL* expression in salivary gland sporozoites. *MAEBL* expression is also detected in blood stages. Amplification with *MSP-1*-specific primers shows *MSP-1* expression in blood stages. *MSP-1* expression is not detected in salivary gland sporozoites. Sizes are given in base pairs (bp). (*C*) Localization of MAEBL by indirect immunofluorescence assay in *P. yoelii* salivary gland sporozoites with antisera against the carboxyl cysteine-rich region. (*D*) Localization of MAEBL by indirect immunofluorescence in *P. yoelii* midgut sporozoites with antisera against the carboxyl cysteine-rich region. Scale bar for *C* and *D* = 1 μm.

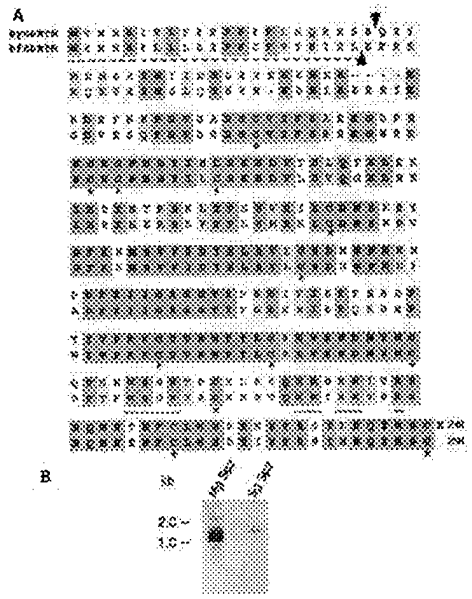

Fig. 4. Alignment of SPATR and expression in sporozoites. (*A*) Comparison of the deduced amino acid sequences of the *P. yoelii* SPATR with the homologue in *P. falciparum* (accession no. C71611). The conserved residues of the altered TSR are underlined with a solid line. The putative signal peptides are underlined with a dashed line. Putative signal peptide cleavage sites are marked with arrowheads (,). Conserved cysteine residues are marked with an asterisk (*). Identical residues are shaded dark gray. Conserved amino acid changes are shaded light gray, and radical changes are not shaded. (*B*) cDNA blot demonstrating *SPATR* expression in midgut sporozoites (Mg Spz) and salivary gland sporozoites (Sg Spz). Sizes are given in kb.

One EST in the data set identified another potential sporozoite invasion ligand, matching a hypothetical ORF on chromosome 2 of *P. falciparum* (PFB0570w; ref. 16). We determined the complete ORF for this *P. yoelii* EST. The predicted protein has a putative cleavable signal peptide predicting that it is secreted (Fig. 4*A*). Significantly, the protein carries a motif with similarity to the thrombospondin type 1 repeat (TSR) (32). We therefore named it SPATR (secreted protein with altered thrombospondin repeat). The most conserved motif of the TSR is present (WSXW), followed by a stretch of basic residues. The central CSXTCG that follows the WSXW motif in a number of the TSR superfamily members (33) is not present in SPATR. Interestingly, this motif is present in the TSR of CS but it is not important for CS binding to the hepatocyte surface (34). The *P. yoelii* and *P. falciparum* SPATR proteins share 63% amino acid sequence identity, including 12 conserved cysteine residues (Fig. 4*A*). The N-terminal intron of *SPATR* is conserved in both species (data not shown). This overall similarity suggests that the proteins are homologous. To confirm *SPATR* transcription, we hybridized a salivary gland and midgut sporozoite cDNA blot to a *SPATR*-specific probe. *SPATR* cDNA seemed more abundant in the midgut sporozoite preparations (Fig. 4*B*).

One EST showed weak similarity with Pbs48/45, a member of the six-cysteine (6-cys) superfamily (35). A *P. yoelii* contig from the *P. yoelii* genome project that matched this EST showed a single ORF of 1,440 bp coding for a predicted mature 52-kDa protein. Search of the *P. falciparum* genome database identified a putative homologue that shared 40% amino acid sequence identity with the *P. yoelii* protein (Fig. 5*A*). Both predicted proteins have consensus amino terminal cleavable signal peptides followed by two tandem 6-cys domains. A carboxyl-terminal hydrophobic domain indicated that the proteins could be membrane-anchored by a glycosylphasphatidylinositol linkage. The presence of the 6-cys domain and the overall structure clearly identified the proteins as new members of the 6-cys superfamily. According to the nomenclature of this superfamily by predicted molecular mass of the mature protein, we named the proteins Py52 and Pf52. To confirm *Py52* expression, we hybridized a salivary gland and midgut sporozoite cDNA blot to a *Py52* specific probe. *Py52* cDNA seemed more abundant in the midgut sporozoite preparations (Fig. 5*B*).

Finally, it is noteworthy that none of our ESTs resulted in significant matches with sporozoite-threonine asparagine-rich protein and liver stage antigen-3, proteins that have been described in *P. falciparum* sporozoites (12, 13).

Discussion

The nearly complete genome sequence of *P. falciparum* is now available, and its annotation will be concluded in the near future (36). It has been estimated that the 25–30 megabase genome harbors about 6,000 expressed genes. In addition, a 2× sequence coverage of the *P. yoelii* genome has very recently been completed and made publicly available (www.tigr.org). Malaria parasites occur in a number of different life cycle stages, making it a challenging task to determine which subset of the 6,000 genes is represented in the transcriptome of each stage. Microarrays will be the method of choice for expression analysis in asexual and sexual blood stage parasites where the acquisition of sufficient RNA is not a limitation. Although whole genome microarrays are not yet available, partial arrays from mung bean genomic libraries (37) or blood stage cDNA libraries (38) have been used successfully to study gene expression in blood stages. However, microarray analysis of gene expression in ookinetes, early oocysts, sporozoites, and EEF of mammalian Plasmodia will be difficult because large quantities of these stages are not available.

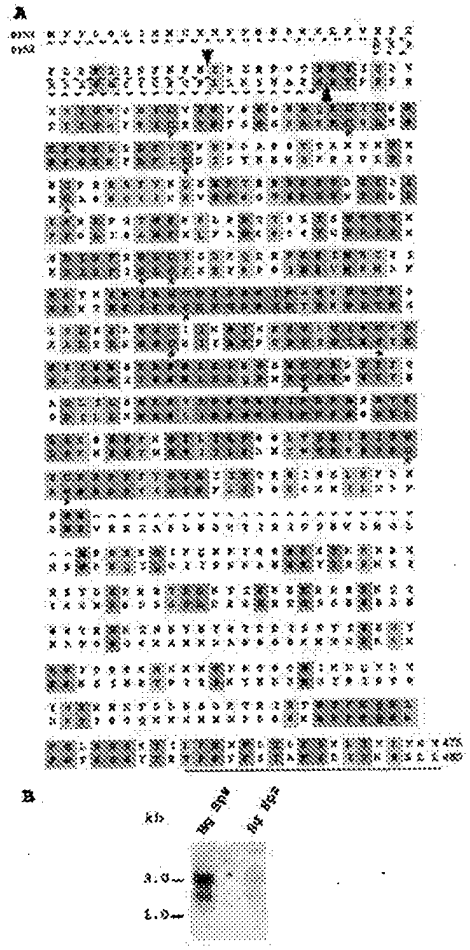

Fig. 5. Alignment of P52 and expression in sporozoites. (*A*) Comparison of the deduced amino acid sequences of the *P. yoelii*, Py52, with the homologue in *P. falciparum*, Pf52. The putative signal peptides are underlined with a dashed line. Putative signal peptide cleavage sites are marked with arrowheads (,). Conserved cysteine residues of the tandem 6-cys motifs are marked with an asterisk (*). The carboxyl-terminal hydrophobic putative membrane anchor is underlined with a solid line. Identical residues are shaded dark gray. Conserved amino acid changes are shaded light gray, and radical changes are not shaded. (*B*) cDNA blot demonstrating *Py52* expression in midgut sporozoites (Mg Spz) and salivary gland sporozoites (Sg Spz). Sizes are given in kb.

Herein, we have described a survey of genes expressed in the infectious *Plasmodium* salivary gland sporozoite. We have demonstrated that, with a PCR-based amplification of the transcriptome, it is possible to obtain enough cDNA to construct a library for EST sequence acquisition. CS and SSP2/TRAP are highly expressed in the salivary gland sporozoites. On the basis of Western blot analysis of salivary gland sporozoites, CS is more abundant than SSP2/TRAP (data not shown), and this result is in agreement with the number of ESTs for CS (33 ESTs) and SSP2/TRAP (13 ESTs). We do not know whether the low number of ribosomal protein ESTs in the cDNA data set reflects true abundance of transcripts for those proteins in the sporozoite. PCR amplification of cDNA before cloning and sequencing could have biased the representation. Yet, it is possible that the bulk of proteins of the translation machinery are synthesized in the developing oocyst or in midgut sporozoites. The EST data set gives unprecedented insight into sporozoite gene expression, opening up new avenues of exploration. Expression of chorismate synthase in sporozoites is one example. The shikimate pathway was shown to be functional in blood stage *Plasmodium*, and the herbicide glyphosate had a clear inhibitory effect on parasite growth (15). If the shikimate pathway is also operational in sporozoites and EEF, inhibitory drugs (39) could be used to eliminate the preerythrocytic stages, avoiding progression to the blood stage and therefore disease.

The presence of MAEBL in the sporozoite stage raises interesting questions about its function. Binding of MAEBL to erythrocytes suggested that it had a role in merozoite red blood cell invasion (14). It will be worthwhile to investigate whether MAEBL also has a role in mosquito salivary gland and hepatocyte invasion, and therefore acts as a multifunctional parasite ligand in the merozoite and sporozoite stages. Regardless, its dual expression could make MAEBL the target of an inhibitory immune response against erythrocytic and preerythrocytic stages.

We show here that sporozoites express *SPATR*, coding for a putative secreted protein with a degenerate TSR. The CS protein and SSP2/TRAP each carry a TSR, and both proteins have demonstrated roles in sporozoite motility, host cell attachment, and invasion (34, 40–42). TSRs are also present in CS/TRAP-related protein (43), a protein essential for ookinete motility and host cell invasion (44–46).

The 6-cys motif defines a superfamily of proteins that seems to be restricted to the genus *Plasmodium* (35). Where studied, expression of members of this family was restricted to sexual erythrocytic stages. Recently, targeted gene disruption of *P48/45* identified the protein as a male gamete fertility factor (47). We have identified *Py52* and *Pf52* as genes coding for new members of the 6-cys family. *Py52* is expressed in sporozoites, and, like *SPATR*, *Py52* was expressed at higher level in midgut sporozoites than in salivary gland sporozoites. These expression patterns contrast with expression patterns of *SSP2/TRAP* and *CS*, which appeared equally abundant in both sporozoite stages (data not shown). Although we have not yet analyzed SPATR and Py52 protein expression, it is tempting to speculate, based on transcript level, that both proteins may have a role in sporozoite invasion of the mosquito salivary glands.

We have presented and discussed here only an initial analysis of the EST data set and further characterized a few selected examples with emphasis on putative sporozoite ligands for host cell attachment and invasion. A detailed analysis of all ESTs is beyond the scope of this first description. The amount of redundancy present in the EST data set is relatively low. It is therefore likely that the generation of more sequence data will identify novel sporozoite-expressed genes. However, many ESTs do not have significant database matches, and a number of ESTs produce matches with proteins of unknown function. A comprehensive expression analysis will determine which subset of the identified genes is exclusively expressed in the sporozoite stages. Sporozoite-specific genes are amenable to functional genetic analysis because loss-of-function mutants can be isolated and analyzed (48), a tool not yet available for genes essential in the asexual erythrocytic cycle (49). All told, we can now generate more of the urgently needed information about the sporozoite stage, a stage of the complex malaria life cycle that has so far eluded comprehensive experimental study.

Note Added in Proof. Recently, 1,117 additional ESTs were generated. These ESTs are not included in the analysis presented here. The additional ESTs have been deposited in the GenBank dbEST database (accession nos. BG603043–BG604160) and are also available through the *P. yoelii* gene index (http://www.tigr.org/tdb/pygi/).

MICROBIOLOGY

We thank Tirza Doniger at the New York University School of Medicine Research Computing Resource for bioinformatics support. This work was supported by National Institutes of Health Grant AI-47102, the United Nations Development Program/World Bank/World Health Organization Special Program for Research and Training in Tropical Diseases (TDR), the Naval Medical Research Center Work Units 61102AA0101BFX and 611102A0101BCX, and a U.S. Army Medical Research and Material Command Contract (DAMD17-98-2-8005). S.H.I.K. is a recipient of the B. Levine fellowship in malaria vaccinology. We thank the scientists and funding agencies comprising the international Malaria Genome Project for making sequence data from the genome of *P. falciparum* (3D7) public prior to publication of the completed sequence. The Sanger Centre (Hinxton, U.K.) provided sequence for chromosomes 1, 3-9, and 13, with financial support from the Wellcome Trust. A consortium composed of the Institute for Genome Research, along with the Naval Medical Research Center (Silver Spring, MD) sequenced chromosomes 2, 10, 11, and 14, with support from the National Institute of Allergy and Infectious Diseases/National Institutes of Health, the Burroughs Wellcome Fund, and the Department of Defense. The Stanford Genome Technology Center sequenced chromosome 12, with support from the Burroughs Wellcome Fund. The Plasmodium Genome Database is a collaborative effort of investigators at the University of Pennsylvania and Monash University (Melbourne, Australia) supported by the Burroughs Wellcome Fund.

1. Holder, A. A. (1996) in *Malaria Vaccine Development: A Multi-Immune Response Approach*, ed. Hoffman, S. L. (Am. Soc. Microbiol., Washington, DC), pp. 35–75.
2. Nussenzweig, V. & Nussenzweig, R. S. (1989) *Adv. Immunol.* **45,** 283–334.
3. Nussenzweig, R. S. & Nussenzweig, V. (1989) *Rev. Infect. Dis.* **11,** S579–S585.
4. Schofield, L., Villaquiran, J., Ferreira, A., Schellekens, H., Nussenzweig, R. S. & Nussenzweig, V. (1987) *Nature (London)* **330,** 664–666.
5. Schofield, L., Ferreira, A., Altszuler, R., Nussenzweig, V. & Nussenzweig, R. S. (1987) *J. Immunol.* **139,** 2020–2025.
6. Ferreira, A., Schofield, L., Enea, V., Schellekens, H., van der Meide, P., Collins, W. E., Nussenzweig, R. S. & Nussenzweig, V. (1986) *Science* **232,** 881–884.
7. Sinnis, P. & Nussenzweig, V. (1996) in *Malaria Vaccine Development: A Multi-Immune Response Approach*, ed. Hoffman, S. L. (Am. Soc. Microbiol., Washington, DC), pp. 15–33.
8. Hoffman, S. L., Franke, E. D., Hollingdale, M. R. & Druilhe, P. (1996) in *Malaria Vaccine Development: A Multi-Immune Response Approach*, ed. Hoffman, S. L. (Am. Soc. Microbiol., Washington, DC), pp. 35–75.
9. Charoenvit, Y., Leef, M. F., Yuan, L. F., Sedegah, M. & Beaudoin, R. L. (1987) *Infect. Immun.* **55,** 604–608.
10. Rogers, W. O., Malik, A., Mellouk, S., Nakamura, K., Rogers, M. D., Szarfman, A., Gordon, D. M., Nussler, A. K., Aikawa, M. & Hoffman, S. L. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 9176–9180.
11. Robson, K. J., Hall, J. R., Jennings, M. W., Harris, T. J., Marsh, K., Newbold, C. I., Tate, V. E. & Weatherall, D. J. (1988) *Nature (London)* **335,** 79–82.
12. Fidock, D. A., Bottius, E., Brahimi, K., Moelans, I. M. D., Aikawa, M., Konings, R. N., Certa, U., Olafsson, P., Kaidoh, T., Asavanich, A., *et al.* (1994) *Mol. Biochem. Parasitol.* **64,** 219–232.
13. Daubersies, P., Thomas, A. W., Millet, P., Brahimi, K., Langermans, J. A. M., Ollomo, B., Mohamed, L. B., Slierendregt, B., Eling, W., Van Belkum, A., *et al.* (2000) *Nat. Med.* **6,** 1258–1263.
14. Kappe, S. H. I., Noe, A. R., Fraser, T. S., Blair, P. L. & Adams, J. H. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 1230–1235.
15. Roberts, F., Roberts, C. W., Johnson, J. J., Kyle, D. E., Krell, T., Coggins, J. R., Coombs, G. H., Milhous, W. K., Tzipori, S., Ferguson, D. J. P., Chakrabarti, D. & McLeod, R. (1998) *Nature (London)* **393,** 801–805.
16. Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., *et al.* (1998) *Science* **282,** 1126–1132.
17. Matuschewski, K., Mota, M. M., Pinder, J. C., Nussenzweig, V. & Kappe, S. H. I. (2001) *Mol. Biochem. Parasitol.* **112,** 157–161.
18. Wan, K. L., Blackwell, J. M. & Ajioka, J. W. (1996) *Mol. Biochem. Parasitol.* **75,** 179–186.
19. Ajioka, J. W., Boothroyd, J. C., Brunk, B. P., Hehl, A., Hillier, L., Manger, I. D., Marra, M., Overton, G. C., Roos, D. S., Wan, K. L., *et al.* (1998) *Genome Res.* **8,** 18–28.
20. Strong, W. B. & Nelson, R. G. (2000) *Mol. Biochem. Parasitol.* **107,** 1–32.
21. Chakrabarti, D., Reddy, G. R., Dame, J. B., Almira, E. C., Laipis, P. J., Ferl, R. J., Yang, T. P., Rowe, T. C. & Schuster, S. M. (1994) *Mol. Biochem. Parasitol.* **66,** 97–104.
22. Pinder, J. C., Fowler, R. E., Dluzewski, A. R., Bannister, L. H., Lavin, F. M., Mitchell, G. H., Wilson, R. J. & Gratzer, W. B. (1998) *J. Cell. Sci.* **111,** 1831–1839.
23. Margos, G., Siden-Kiamos, I., Fowler, R. E., Gillman, T. R., Spaccapelo, R., Lycett, G., Vlachou, D., Papagiannakis, G., Eling, W. M., Mitchell, G. H. & Louis, C. (2000) *Mol. Biochem. Parasitol.* **111,** 465–469.
24. Heintzelman, M. B. & Schwartzman, J. D. (1997) *J. Mol. Biol.* **271,** 139–146.
25. Heintzelman, M. B. & Schwartzman, J. D. (1999) *Cell Motil. Cytoskeleton* **44,** 58–67.
26. Bonhomme, A., Bouchot, A., Pezzella, N., Gomez, J., Le Moal, H. & Pinon, J. M. (1999) *FEMS Microbiol. Rev.* **23,** 551–561.
27. Kieschnick, H., Wakefield, T., Narducci, C. A. & Beckers C. (2001) *J. Biol. Chem.* **276,** 12369–12377.
28. Cassaing, S., Fauvel, J., Bessieres, M. H., Guy, S., Seguela, J. P. & Chap, H. (2000) *Int. J. Parasitol.* **30,** 1137–1142.
29. Mota, M. M., Pradel, G., Vanderberg, J. P., Hafalla, J. C. R., Frevert, U., Nussenzweig, R. S., Nussenzweig, V. & Rodriguez, A. (2001) *Science* **291,** 141–144.
30. Kappe, S. H. I., Curley, G. P., Noe, A. R., Dalton, J. P. & Adams, J. H. (1997) *Mol. Biochem. Parasitol.* **89,** 137–148.
31. Adams, J. H., Sim, B. K. L., Dolan, S. A., Fang, X., Kaslow, D. C. & Miller, L. H. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 7085–7089.
32. Lawler, J. & Hynes, R. O. (1986) *J. Cell Biol.* **103,** 1635–1648.
33. Adams, J. C. & Tucker, R. P. (2000) *Dev. Dyn.* **218,** 280–299.
34. Gantt, S. M., Clavijo, P., Bai, X., Esko, J. D. & Sinnis, P. (1997) *J. Biol. Chem.* **272,** 19205–19213.
35. Templeton, T. J. & Kaslow, D. C. (1999) *Mol. Biochem. Parasitol.* **101,** 223–227.
36. Carucci, D. J. & Hoffman, S. L. (2000) *Nat. Med.* **6,** 1–6.
37. Hayward, R. E., Derisi, J. L., Alfadhli, S., Kaslow, D. C., Brown, P. O. & Rathod, P. K. (2000) *Mol. Microbiol.* **35,** 6–14.
38. Mamoun, C. B., Gluzman, I. Y., Hott, C., MacMillan, S. K., Amarakone, A. S., Anderson, D. L., Carlton, J. M.-R., Dame, J. B., Chakrabarti, D., Martin, R. K., *et al.* (2001) *Mol. Microbiol.* **39,** 26–36.
39. McConkey, G. A. (1999) *Antimicrob. Agents Chemother.* **43,** 175–177.
40. Sinnis, P. (1996) *Infect. Agents Dis.* **5,** 182–189.
41. Sultan, A. A., Thathy, V., Frevert, U., Robson, K. J., Crisanti, A., Nussenzweig, V., Nussenzweig, R. S. & Ménard, R. (1997) *Cell* **90,** 511–522.
42. Kappe, S., Bruderer, T., Gantt, S., Fujioka, H., Nussenzweig, V. & Ménard, R. (1999) *J. Cell Biol.* **147,** 937–944.
43. Trottein, F., Triglia, T. & Cowman, A. F. (1995) *Mol. Biochem. Parasitol.* **74,** 129–141.
44. Dessens, J. T., Beetsma, A. L., Dimopoulos, G., Wengelnik, K., Crisanti, A., Kafatos, F. C. & Sinden, R. E. (1999) *EMBO J.* **18,** 6221–6227.
45. Yuda, M., Sakaida, H. & Chinzei, Y. (1999) *J. Exp. Med.* **190,** 1711–1716.
46. Templeton, T. J., Kaslow, D. C. & Fidock, D. A. (2000) *Mol. Microbiol.* **36,** 1–9.
47. van Dijk, M. R., Janse, C. J., Thompson, J., Waters, A. P., Braks, J. A. M., Dodemont, H. J., Stunnenberg, H. G., Van Gemert, G.-J., Sauerwein, R. W. & Eling, W. (2001) *Cell* **104,** 153–164.
48. Ménard, R. & Janse, C. (1997) in *Methods: A Companion to Methods in Enzymology—Analysis of Apicomplexan Parasites* (Academic, Orlando, FL), Vol. 13, pp. 148–157.
49. De Koning-Ward, T. F., Janse, C. J. & Waters, A. P. (2000) *Annu. Rev. Microbiol.* **54,** 157–185.